ENOD2 GENE REGULATORY REGION

Field of the Invention

5   The field of this invention is the area of plant molecular biology in general, and relates in particular to plant genetic engineering by recombinant DNA technology. This invention specifically relates to a soybean early nodulin gene regulatory region which regulates downstream

10   gene expression in a tissue-specific fashion in the developing soybean root nodule after inoculation with Bradyrhizobium japonicum.

Background of the Invention

        Nitrogen-fixing root nodules of leguminous plants are

15   formed as the result of root infection by rhizobia and subsequent development of a symbiosis between bacteria and plant. The development of the symbiosis is dependent on specific recognition between plant and bacterium, and it requires genetic information from both the plant and the

20   bacteria.

        Nodule development displays variation among legumes. Two different types of nodules are recognized, determinant

and indeterminant. The nodules of soybean (<u>Glycine</u> <u>max</u>), for example, are determinant and spherical in shape. In contrast, the nodules of alfalfa (<u>Medicago</u> spp.), clover (<u>Trifolium</u> spp.), and pea (<u>Pisum</u> <u>sativum</u>) are indeterminant and elongated in shape. These nodules are also anatomically and metabolically distinct, reflecting differences in the process of nodule development which may be attributable to genetic differences between legumes as well as between the different species of <u>Rhizobium</u> which infect them.

In a description of nodule development, Vincent (J.M. Vincent (1980) in <u>Nitrogen</u> <u>Fixation</u>, eds. W.E. Newton and W.H. Orme-Johnson (University Park Press, Baltimore, Maryland, Vol. 2, pp. 103-131) distinguishes between three different stages of nodule formation: preinfection, infection and nodule formation, and nodule function. In the preinfection stage, the <u>Rhizobium</u> cells recognize their host plants and attach to root hairs, an event which is followed by root hair curling. In the next stage, the bacteria enter the roots via infection threads while some cortical cells dedifferentiate to form meristem. The infection threads grow toward the meristematic cells. Bacteria are released into the cytoplasm of about half of these cells, and subsequently the bacterial cells develop into bacteroids. In the final stage further differentia-

tion of nodule cells occurs leading up to a nitrogen-fixing nodule.

Nodule-specific proteins, which are only expressed in root nodules, are likely to be associated with the infection process, nodule development, and symbiotic nitrogen fixation. Both proteins of plant origin (nodulins) and of bacteroid origin (bacteroidins) are found in nodules. Nodule-specific proteins have been identified in root preparations of soybean infected with Bradyrhizobium japonicum (R.P. Legocki and D.P.S. Verma (1980), supra) and pea (Pisum sativum) infected with Rhizobium leguminosarum PRE (T. Bisseling et al. (1983) EMBO J. 2:961-966). In each case, a nodule-specific antiserum was used to identify the nodule proteins by immunoprecipitation. Each of these antisera was produced by titration of an antiserum raised against soluble nodule proteins with a root preparation from uninfected plants. The drawbacks to these studies are that the plant or bacterial origin of the nodule-specific proteins could not be established and that the antigenicity of each protein affects the immunological analysis.

In soybean, the in vitro translation products of root nodule polysomes were analyzed with nodule-specific antiserum. Control experiments showed that bacterial RNA was not translated in the in vitro system. At least 18-20

3

host plant-derived polypeptides were identified having molecular weights in the range of 18-20 kd. These proteins were absent from uninfected roots, bacteroids and free-living B. japonicum (R.P. Legocki and D.P.S. Verma (1980) Cell 20:153-163). In addition, bacteroids were isolated and incubated with ^{35}S methionine to label bacteroid proteins. Two polypeptides cross-reacted with nodule-specific antiserum. The bacteroid excreted polypeptides had molecular weights of about 11 kd (R.C. van den Bos et al. (1978) J. Gen. Microbiol. 109:131-139). Approximately 20 nodule-specific proteins were identified in pea root protein extracts by probing Western protein blots with nodule-specific antiserum. The proteins de-tected ranged in molecular weight from 15 to 120 kd; however the origin of these proteins was not determined. In these experiments the in vivo nodule proteins were identified (T. Bisseling et al. (1983), supra), while the soybean study analyzed potentially truncated products of in vitro translation.

Verma and co-workers have also isolated soybean nodulin cDNA clones (F. Fuller et al. (1983) Proc. Natl. Acad. Sci. USA 80:2594-2598). Those clones were used to hybrid select NOD mRNAs from nodule RNA preparations; mRNAs of about 1150, 770, and 3150 nucleotides in length yielded in vitro translation products of 27, 24, and 100 kDa, respectively. Two additional clones, which shared some

homology with each other, hybrid selected mRNAs of 1600 and 1100 nucleotides in length with in vitro translation products of ·23.5 and 24.5 kDa, respectively (F. Fuller and D.P.S. Verma (1984) Plant Mol. Biol. 3:21-28) were

5   identified.

Nodule mRNA from different stages of developing pea nodules was studied by in vitro translation of the RNA followed by separation of translation products by two

10   dimensional gel electrophoresis. Twenty-one nodule-specific proteins were found, with molecular weights ranging from 15 to 80 kDa (F. Govers et al. (1985) EMBO J. 4:861-867).

Among the nodulins with known functions are

15   leghemoglobin (C.A. Appleby (1984) Ann. Rev. Plant Physiol. 35:443-478), a nodule-specific glutamine synthetase (J.V. Callimore et al. (1983) Planta 157:245-253), and a nodule-specific form of uricase (M. Bergmann et al. (1983) EMBO J. 2:2333-2339). The functions of most nodulins have

20   not been defined. Nodulins may have specific functions in the formation of nodule tissue after the dedifferentiation and proliferation of cortical cells, in the ·transport of substrates to the bacteroids, in the assimilation of ammonia excreted by the bacteroids, or in the senescence of

25   nodule tissue.

A cDNA library prepared from mature (21 day) soybean root nodules infected with <u>Bradyrhizobium japonicum</u> has been analyzed for copies of mRNA transcripts of early (7 day) nodulin genes (Franssen <u>et</u> <u>al</u>. (1987) Proc. Natl. Acad. Sci. USA <u>84</u>:4495-4499). These genes are expressed while the nodule structure is being formed. pEnod2, the cDNA clone whose insert encodes nodulin-75 (N-75) was sequenced. The 998 bp insert includes a short poly(A) tail, and encodes a proline-rich protein. Nodule mRNA of about 1200 nucleotides in length was hybrid-selected and translated <u>in</u> <u>vitro</u> to give two polypeptides each with an M_r of about 75 kDa. The coding capacity of the mRNAs is significantly less than 75 kDa, but proline-rich proteins, such as collagen, are known to have anomalous behavior on polyacrylamide gels (J.W. Freytag <u>et</u> <u>al</u>. (1979) Biochemistry <u>18</u>:4761-4768). N-75 expression was first detected at day 7 of nodule development, when nodule meristem emerges through the root epidermis with apparent expression increasing up to about day 13. Expression was observed in <u>R. fredii</u>-induced ineffective nodules without infection threads or bacteroids, so N-75 is likely to be involved in nodule morphogenesis rather than in the infection process <u>per</u> <u>se</u> (H. Franssen <u>et</u> <u>al</u>. (1987) Proc. Natl. Acad. Sci. USA <u>84</u>:4495-4599).

There is a growing understanding of the DNA sequence elements which control gene expression. The following

discussion applies to plant genes which are transcribed by polymerase II. There are known sequences which direct the initiation of mRNA synthesis, those which control transcription in response to environmental stimuli, those

5   which modulate the level of transcription and there are those which regulate gene expression in a tissue-specific fashion.

Promoters are the portions of DNA sequence at the beginnings of genes, which contain the signals for RNA

10   polymerase to begin transcription so that protein synthesis can then proceed. Eukaryotic promoters are complex, and are comprised of components which include a TATA box consensus sequence in the vicinity of -30 relative to the transcription start site (+1) (R. Breathnach and P. Chambon

15   (1981) Ann. Rev. Biochem. 50:349-383; C. Kuhlemeier et al. (1987) Ann. Rev. Plant Physiol. 38:221-257). In plants there may be substituted for the CAAT box a consensus sequence which J. Messing et al. (1983) in Genetic Engineering of Plants, T. Kosuge, C. Meredith, and A.

20   Hollaender, eds., have termed the AGGA box, positioned a similar distance from the cap site (+1). Other sequences in the 5' regions of genes are known which regulate the expression of downstream genes. There are sequences which participate in the response to environmental conditions,

25   such as illumination, nutrient availability, hyperthermia, anaerobiosis, or the presence of heavy metals. There are

also signals which control gene expression during development, or in a tissue-specific fashion. Promoters are usually positioned 5' to, or upstream of, the start of the coding region of the corresponding gene, and the DNA tract containing the promoter sequences and the ancillary promoter-associated sequences affecting regulation or the absolute levels of transcription may be comprised of less than 100 bp or as much as 1 kbp.

As defined by G. Khoury and P. Gruss (1983) Cell 22:313-314, an enhancer is one of a set of eukaryotic promoter-associated elements that appears to increase transcriptional efficiency in a manner relatively independent of position and orientation with respect to the nearby gene. The prototype enhancer is found in the animal virus SV40. Generally animal or animal virus enhancers can function over a distance as much as 1 kbp 5', in either orientation, and can act 5' or 3' to the gene. The identifying sequence motif (5'-GTGGAAA(orTTT)G-3') is generally reiterated. There have been sequences identified in or adjacent to plant genes which have homology to the core consensus sequence of the SV40 enhancer, but the functional significance of these sequences in plants has not been determined.

There are also reports of enhancer-like elements 5' to certain constitutive and inducible genes of plants. J.

8

Odell _et al_. (1985), Nature 313:810-812, describe a stretch of about 100 bp 5' to the start site of the CaMV 35S transcript which is necessary for increasing the level of expression of a reporter gene in chimeric constructions.

5   Two different transcription activating elements which can function in plants are derived from the 780 gene and the ocs gene of _Agrobacterium tumefaciens_ T-DNA (W. Bruce and W. Gurley (1987) Mol. Cell. Biol. 7:59-67; J. Ellis _et al_. (1987) EMBO J. 6:11-16). Regulated enhancer-like elements

10   include those believed to mediate tissue-specific expression and response to illumination (M. Timko _et al_. (1985) Nature 318:579-582; H. Kaulen _et al_. (1986) EMBO J. 5:1-8; J. Simpson _et al_. (1985) EMBO J. 4:2723-2729; J. Simpson _et al_. (1986) Nature 323:551-554; R. Fluhr _et al_.

15   (1986) Science 232:1106-1112).

The molecular mechanisms which regulate the expression of nodulin genes are not yet defined. V.P. Mauro _et al_. (1985) Nucleic Acids Res. 13:239-249, have analyzed the 5' flanking sequences of three nodulin genes

20   of soybean for conserved DNA sequence motifs. They found three conserved sequence motifs: consensus sequence a 5'-GTTTCCCT-3', consensus sequence b 5'-GGTAGTG-3', and consensus sequence c 5'-TCTGGGAAA-3'. Whether these sequences function in the regulation of the nodulin genes

25   is not known, and if they do, the stimuli which elicit expression are not known. The molecular mechanisms

controlling the expression of Enod2 genes in soybean are also not known, but F. Govers et al. (2986) Nature 323:564-566, have shown that in developing pea root nodules, Rhizobium leguminosarum nod genes or adjacent genes carried on a 10 kb region of the Sym plasmid are involved in inducing an early nodulin gene which is homologous to the Enod2 gene of soybean.

Jensen et al. (1986) Nature 321:669-674, transformed the wild legume Lotus corniculatus with a Leghemoglobin-CAT chimeric construct. Roots were infected with a strain of Agrobacterium rhizogenes, and transformed plants containing the hybrid gene were obtained. Upon infection with Rhizobium loti, nodules were formed that expressed the introduced CAT gene in a fashion that was correct by all criteria applied.

Summary of the Invention

The work of the present invention describes the isolation and characterization of DNA sequences functional in soybean, each of which regulates the expression of a downstream structural gene during the early stages of soybean root nodule development after inoculation with Bradyrhizobium japonicum. These regulatory regions are unlike previously described regulatory regions from nodulin genes in that they direct expression earlier in nodule

development than other nodulin genes.   These regulatory regions are those of early nodulin genes (Enod2).

5

10

The Enod2 gene, encodes a nodulin-75, a polypeptide with an apparent molecular weight of about 75 kDa expressed during the early stages of nodule development.   The Enod2a regulatory region extends about 1 kb 5' from the start of transcription of the gene.   All the signals required for tissue-specific regulated gene expression are contained within this 1000 bp 5' flanking region.   The Enod2a regulatory region controls the expression of a downstream structural gene in a tissue-specific manner in the cortex of developing soybean nodule early in the nodule development process.

15

20

25

Examples of tissue-specific early nodulin regulatory regions are found in the 5' flanking region of the soybean (Glycine max) Enod2a and Enod2b genes which encode N-75. The Enod2a regulatory region extends about 1 kb 5' from the transcription start of the genes.   The regulatory region contains the nucleotide sequence from Table 1 extending from about nucleotide 520 to about nucleotide 1565.   The Enod2b regulatory region extends about 1 kb 5' from the transcription start of the gene, from about nucleotide 1320 to about nucleotide 2365, as in Table 2.   These regulatory regions direct the expression of a downstream gene in a tissue-specific manner in the developing root nodule.

An additional example of a tissue-specific early nodulin gene regulatory region is the DNA sequence common to the 5' flanking regions of the soybean Enod2a and Enod2b genes. This regulatory element contains DNA sequence as given in Table 1, extending from about nucleotide 1050 to about nucleotide 1565, or given in Table 2, extending from about nucleotide 1850 to about nucleotide 2365. This regulatory region directs the expression of a downstream structural gene in a tissue-specific manner in the developing root nodule.

A primary object of this invention is to enable those skilled in the art to achieve tissue-specific gene expression in soybean root nodules. This object is accomplished by utilizing a DNA sequence, designated an Enod2 regulatory region. This regulatory region directs the expression of a downstream structural gene during the early stages of nodule development. The term Enod2 regulatory region is used generically to designate the nodule specific regulatory region of any Enod2 gene. The Enod2 regulatory region contains promoter sequences as well as promoter-associated sequences which function in the regulation of the expression of a downstream structural gene.

12

The invention provides recombinant DNA molecules which comprise an Enod2 regulatory region and a plant-expressible structural gene, wherein said structural gene is positioned 3' to said regulatory region and under its regulatory control, with the result that the structural gene is expressed in the developing soybean root nodule. In general, any structural gene, including gene fusions, that is expressible in a plant can be employed in the recombinant DNA molecules of the present invention.

The recombinant DNA molecules of the present invention are useful in a method for selectively expressing a desired plant-expressible structural gene in a developing nodule of soybean root. In such a method, a soybean plant is genetically transformed to contain the recombinant DNA molecules of the present invention, which contain an Enod2 regulatory region and the desired structural gene which is positioned such that it is under the regulatory control of the Enod2 regulatory element. A soybean plant thus transformed expresses the desired structural gene in a tissue-specific manner in developing nodules. Specifically, nodule-specific expression of the desired structural gene can be achieved by introducing the recombinant DNA molecules of the present invention into soybean tissue and regenerating a soybean plant from the transformed tissue. The recombinant DNA molecules of the present invention are particularly useful for the

tissue-specific expression of foreign structural genes not naturally occurring in soybean. Transformation of plant cells and tissue with exogenous or foreign DNA and regeneration of plants from transformed cells or tissue can

5   be achieved by any means known to the art.

Brief Description of the Figure

Figure 1 gives a schematic restriction endonuclease map of the soybean Enod2a and Enod2b genes, and the regions which flank them. Schematic diagrams of CHA-6 (containing

10   the Enod2a gene) and CHA-9 (containing the Enod2b gene) are given. The regions sequenced (Tables 1 and 2) of both clones are indicated. The region of approximately 100% homology between the two genomic clones is indicated, as are the regions of the clones homologous to the Enod2 cDNA

15   clone. Restriction endonucleases are labelled as follows: H = HindIII, B = BamHI, S = Sau3A, E = EcoRI.

Detailed Description of the Invention

The following definitions are provided, in order to remove ambiguities to the intent or scope of their usage in

20   the specification and claims.

The Enod2 gene described herein is an early nodulin gene of soybean (Glycine max), which encodes nodulin polypeptides with an apparent molecular weight of about 75 kDa, nodulin 75 (N-75). Two such genes are exemplified by

14

the Enod2a and Enod2b genes which are identified by the DNA sequences given in Tables 1 and 2, respectively.

The Enod2 regulatory region is the DNA sequence 5' and adjacent to the Enod2 coding sequence, which includes pro-
5 moter sequences and promoter-associated sequences and controls tissue-specific expression of the Enod2 genes in soybean. The regulatory region extends about 1 kb upstream from the transcription start site of an Enod2 gene. All the signals required from tissue-specific regulated gene
10 expression are contained in the approximately 1 kb 5' flanking region. Within this stretch of DNA are sequences with homology to the TATA and CAAT consensus sequences of eukaryotic promoters, and the nodulin gene consensus sequences a and c (V.P. Mauro et al. (1985), supra), which
15 are believed to be involved in the regulation of the expression of nod genes expressed later than Enod2 during nodulation. There are also sequence motifs with homology to the SV40 enhancer core consensus sequence which are found in the regulatory region of the soybean Enod2a gene.
20 There may also be other sequence elements which modulate the level of gene expression, which respond to stimuli from the B. japonicum, or which determine the tissue-specific expression in the developing soybean root nodule after inoculation with Bradyrhizobium japonicum. The expression
25 of Enod2 genes controlled by the Enod2 regulatory region is tissue-specific in that it is limited to the cortex of

developing soybean root nodules. The Enod2 regulatory region controls early gene expression in the developing root nodule of soybean with expression beginning at about 7 days after seed planting and inoculation. Expression is induced by contact with soybean nodulating bacteria, such as B. japonicum. Enod2 gene expression also occurs in the ineffective nodules induced by strains of Rhizobium fredii. The Enod2a regulatory region is a DNA sequence which includes promoter sequences and promoter-associated sequences and controls the expression of the soybean Enod2a gene. The Enod2a regulatory region extends about 1 kb upstream from the Enod2a gene transcription start. This region is specifically identified by the DNA sequence in Table 1 from about nucleotide 520 to about nucleotide 1565. The Enod2b regulatory region is a DNA sequence which includes promoter sequences and promoter-associated sequences and controls the expression of the soybean Enod2b gene. The Enod2b regulatory region extends about 1 kb upstream from the Enod2b gene transcription start. This region is specifically identified by the DNA sequence in Table 2 from about nucleotide 1320 to about nucleotide 2365. These regulatory regions direct tissue-specific expression of a downstream structural gene, such that the gene is selectively expressed in the inner cortex of the developing root nodule in soybean. The Enod2 common regulatory region is the DNA sequence extending about 500 bases upstream of the transcription start site of an Enod2

16

gene.   The Enod2 common regulatory region is exemplified by the homologous sequences of Enod2a and Enod2b extending from about nucleotide 1050 to about nucleotide 1565 (Table 1), and about nucleotide 1850 to about nucleotide 2365 (Table 2), respectively.   This common regulatory region controls tissue-specific expression of downstream genes in the cortex of developing soybean root nodules.

5

Expression refers to the transcription and translation of a structural gene so that a polypeptide is made.   Gene expression may be assessed by direct detection of the protein product, by protein electrophoresis or by immunological methods, for example.   Alternatively, expression may be assessed by the detection of the mRNA products of transcription (i.e. by northern hybridizations).   This method is particularly appropriate for the testing of transcriptional regulatory sequences because the effects of processes such as protein degradation are excluded.

10

15

Promoter refers to the DNA sequences at the 5' end of a structural gene which direct the initiation of transcription.   Promoter sequences are necessary, but not always sufficient, to drive the expression of the downstream structural genes.   The promoter itself may be a composite of segments derived from more than one source, naturally occurring or synthetic.   Eukaryotic promoters are

20

25

commonly recognized by the presence of DNA sequence elements homologous to the canonical form 5'-TATAAT-3' (TATA box) about 10-30 bp 5' to the 5' end of the mRNA (cap site, +1). About 30 bp 5' to the TATA box another promoter component sequence is often, but not always, found which is recognized by the presence of DNA sequences homologous to the canonical form 5'-CCAAT-3'. For the purposes of this application, a promoter is considered to extend about 100 bp 5' from the transcription start site. Promoter-associated sequence elements located further upstream from -100, or within the region between -100 and +1, may contribute to, or exert regulatory control and may determine the relative levels of gene expression. DNA sequences associated with regulatory control of gene expression can extend about 1 kb upstream of the transcription start site of a gene. There may also be additional promoter-associated sequences between +1 and the translation start site which contribute to gene regulation either at the transcriptional or the translational level.

Structural gene refers to that portion of a gene comprising a DNA segment coding for a protein, polypeptide or portion thereof, possibly including a ribosome binding site and/or a translational start codon, but lacking at least one component which drives the initiation of transcription. The term can also refer to copies of a structural gene naturally found within a cell but

artificially introduced. The structural gene may encode a protein not normally found in the plant cell in which the gene is introduced, in which case it is termed a foreign structural gene. A foreign structural gene may be derived

5 in whole or in part from a bacterial genome or episome, eukaryotic nuclear or plastid DNA, cDNA, viral DNA, or chemically synthesized DNA. It is further contemplated that a structural gene may contain one or more modifications in either the coding segments or in the

10 untranslated regions which could affect the biological activity or the chemical structure of the expression product, the rate of expression or the manner of expression control. Such modifications include, but are not limited to, insertions, deletions, and substitutions of one or

15 more nucleotides. The structural gene may constitute an uninterrupted coding sequence or it may include one or more introns, bounded by the appropriate splice junctions functional in plants. The structural gene may be a composite of segments derived from one or more sources,

20 naturally occurring or synthetic. That structural gene may also produce a fusion protein. In this application a structural gene is considered to include the polyadenylation signal downstream from the translation termination codon. That polyadenylation signal usually results in the

25 addition of polyadenylic acid tracts to the 3' ends of the precursor mRNAs. It is also known that a canonical polyadenylation signal may cause a cleavage of the

19

transcript and not poly(A) addition per se (C. Montell et al. (1983) Nature 305:600). It is contemplated that the introduction into plant tissue recombinant DNA molecules containing the Enod2 regulatory region/structural gene complex will include constructions wherein the structural gene and the regulatory region are not derived from the same source (heterologous constructions). Such constructions can include those wherein additional copies of a gene naturally expressed in a plant tissue, but not regulated as an Enod2 gene, are transcribed under the regulatory control of the Enod2 regulatory region. It is understood in the art how to combine the requisite functional elements of regulatory regions and structural genes to achieve gene expression in plant tissue.

Regulatory control refers to the modulation of gene expression by sequence elements upstream of the transcription start site. Regulation may result in an on/off switch for transcription, or it may result in variations in the levels of gene expression. To place a structural gene under regulatory control of sequence elements means to place it sufficiently close to such sequence elements, and in a position relative to such sequence elements so that the gene is switched on or off, or so that its level of expression is measurably varied, as is understood by those skilled in the art. There can also be sequence components in the untranslated leader region of

mRNA which contribute to the regulation of gene expression at the translational level.

Chemically synthesized, as related to a sequence of DNA, means that the component nucleotides were assembled in vitro using nonenzymatic means. Manual chemical synthesis of DNA may be accomplished using well established procedures (i.e. M. Caruthers (1983) in Methodology of DNA and RNA Sequencing, Weissman (ed.), Praeger Publishers (New York) Chapter 1), or automated synthesis can be performed using one of a number of commercially available machines. Employing the DNA sequence information provided herein, the Enod2 regulatory regions or portions thereof can be synthesized and these synthetic sequences can then be utilized in the construction of the recombinant DNA molecules of the present invention.

Plant tissue includes differentiated and undifferentiated tissues of plants including, but not limited to, roots, shoots, leaves, pollen, seeds, tumor tissue, such as crown galls, and various forms of aggregations of plant cells in culture, such as embryos and calli. The plant tissue may be in planta or in organ, tissue, or cell culture.

Homology as used herein, refers to identity of nucleotide sequences. The extent of homology between DNA

sequences can be empirically determined in DNA hybridization experiments, such as those described in B. Hames and S. Higgins (1985) *Nucleic* *Acid* *Hybridization*, IRL Press, Oxford, UK.

5   pEnod2 was isolated from a cDNA library prepared with 21-day-old soybean root nodule RNA, using RNA from 10-day-old nodules as a probe. Thus, pEnod2 represents an early nodulin cDNA clone. The early nodulin encoded by pEnod2 was identified by hybrid-selecting nodule mRNA and

10   translating *in vitro*. Two polypeptides, with apparent M_rs of 75000, were found and were each called N-75. The mRNAs homologous to pEnod2 were only about 1200 nucleotides long, with the capacity to encode a protein of at most about 45 kDa. Therefore the soybean-specific insert of pEnod2 was

15   sequenced and the amino acid sequence of N-75 was deduced. Two ORFs of similar size were found (labelled ORF1 and ORF2 on Tables 1 and 2), one with about 20 methionines and the other a proline-rich sequence, with a repeating heptameric sequence. Because of the anomalous migration on

20   SDS-polyacrylamide gels and because of the labelling patterns the two N-75s, it was concluded that the proline-rich coding sequence (ORF1) was that of N-75. It is believed that N-75 is involved in nodule morphogenesis because of its proline content and because of the pattern

25   of expression in the developing nodule. N-75 appears at about day 7 after sowing and inoculation, and increases

through day 13; mRNA continues to be present at least through day 21. N-75 is also produced in the developing ineffective nodule of soybean inoculated with _Rhizobium fredii_ USDA257. That leads to the conclusion that typical

5   nodule structure with successful infection of the root by rhizobia is not absolutely required for Enod2 expression.

Hybridization studies have shown that there are Enod2 cDNA homologous sequences in _Pisum sativum_, _Vicia sativa_, _Parasponia_, and alfalfa. In pea, the _nod_ genes or genes

10 adjacent to the _nod_ genes of _Rhizobium leguminosarum_ are known to be involved in the expression of the Enod2-homologous gene (F. Govers _et al_. (1986) Nature _323_:564-566).

Two soybean genomic clones corresponding to pEnod2

15 have been isolated and the DNA sequences of the coding and flanking regions have been determined (Tables 1 and 2). The genes, termed Enod2a and Enod2b, are essentially homologous from about 600 bp 5' to the ATG translation start codon through the coding region, which is not

20 interrupted by introns, and through some 500 bp of 3' flanking sequence. Comparison of the genomic clones with the Enod2 cDNA sequence indicates that one or both of these genes are expressed in the developing root nodule. S1 mapping of the transcription start site led to the

25 conclusion that the Enod2a start site is at nucleotide 1543

23

± 20 as shown in Table 1, and the Enod2b start site is deduced to be similarly located at about nucleotide 2350, as shown in Table 2.

The DNA sequence of the Enod2a gene was analyzed for motifs which are believed to function in transcriptional regulation. A sequence with homology to the canonical TATA box sequence was found at about nucleotide 1490, upstream from the transcription start site (between 1523 and 1563). A CAAT box-homologous sequence was found at about 1478. There were two motifs with homology to the NOD consensus sequence a at about 1450 and 1460, and one sequence motif with homology to the NOD consensus sequence c at about 1550, near the cap site. Within about 1 kb of 5' flanking sequence, there are 5 sequences with homology (up to 2 mismatches) to the enhancer sequence 5'-GTGGTTGT-3', at about 567, 979, 1027, 1377, and 1404.

The functionality of any DNA sequences within the Enod2 regulatory region can be tested by those skilled in the art of plant molecular biology. It will be understood that there may be minor variations within sequences utilized or disclosed in the present application. It is well known in the art that some DNA sequences within a larger stretch of sequence are more important than others in determining functionality. A skilled artisan can test allowable variations in sequence by mutagenic techniques

24

which include, but are not limited to, those discussed by
D. Shortle et al. (1981) Ann. Rev. Genet. 15:265; M. Smith
(1985) ibid. 19:423; D. Botstein and D. Shortle (1985)
Science 229:1193; S. McKnight and R. Kingsbury (1982)
Science 217,316; R. Myers et al. (1986) Science 232:613.
It is also known how to generate and analyze deletions of
varying lengths (e.g. T. Maniatis et al. (1982) Molecular
Cloning, Cold Spring Harbor Laboratory, Cold Spring Harbor,
New York). These variations and others can be determined
by standard techniques to enable those of ordinary skill in
the art to manipulate and bring into utility the functional
units of promoter element and structural genes.

Production of genetically modified plant tissue
expressing a structural gene under the transcriptional
control of an Enod2 gene regulatory region functional in
soybean, or in another species of plant, combines the
specific teachings of the present disclosure with a variety
of techniques and expedients known in the art. In most
instances, alternative expedients exist for each stage of
the overall process. The choice of expedients depends on
such variables as the choice of the vector system for the
introduction and stable maintenance of the expression
complex, the plant species to be modified and the desired
regeneration strategy, and the particular structural gene
to be used. Those of ordinary skill are able to select and
use appropriate alternative process steps to achieve a

desired result. For instance, although the ultimate starting point for obtaining the plant regulatory element giving tissue-specific expression during the early stages of soybean root nodule development is the Enod2a or Enod2b

5   genes of _Glycine_ _max_, exemplified in the present application, homologous DNA sequences of other soybean Enod2 genes, or from different sources, could be substituted as long as the appropriate modifications are made to the procedures for manipulating the DNA carrying the Enod2

10   regulatory region, and provided it is known that the regulation afforded by the alternative sequences is equivalent to that determined by that of the soybean Enod2 gene regulatory region. Homologs Enod2 of structural genes or of other sequences may be identified by the ability of

15   their nucleic acids to cross-hybridize under conditions of appropriate stringency as is well understood in the art.

A principal feature of the present invention is a recombinant DNA molecule having a plant-expressible gene

20   whose expression is controlled by the Enod2 regulatory region of soybean. The expression complex comprises the promoter and promoter-associated sequences of the soybean Enod2 regulatory region and a structural gene expressible in a plant. The regulatory region and the structural gene

25   must be correctly positioned and oriented relative to one another such that the promoter sequences and the promoter-associated regulatory sequence can activate

26

transcription of the structural gene in a tissue-specific fashion in the developing root nodule. To be controlled by the Enod2 regulatory region, the structural gene must be inserted on the 3' side of the regulatory region so that

5   the 5' end of the gene is adjacent to the 3' end of the regulatory region. A polyadenylation signal must be located in the correct orientation downstream from the 3' end of the coding sequence. Another consideration is the distance between the functional elements of the expression

10   complex. Substantial variation appears to exist with regard to these distances; therefore, the distance require- ments are best described in terms of functionality. As a first approximation, reasonable operability can be obtained when the distances between functional elements are similar

15   to those in the genes from which they were derived. The distance between the promoter sequences and the 5' end of the structural gene, or between the upstream promoter-associated sequence elements which are responsible for regulatory control and other components in the

20   construction can be varied, and thus one can achieve variations in the levels of expression of the downstream structural gene. In the case of constructions yielding fusion proteins, an additional requirement is that the ligation of the two genes or fragments thereof must be such

25   that the two coding sequences are in the same reading frame, a requirement well understood in the art. An excep- tion to this requirement exists in the case where an intron

27

separates the coding sequence derived from one gene from the coding sequence of the other. In that case, the coding sequences must be bounded by compatible splice sites, and the intron splice sites must be positioned so that the correct reading frame for both genes is established in the fusion after the introns are removed by post-transcriptional processing. It is generally understood in the art how to achieve gene expression in plants, and the skilled artisan will ensure that all necessary requirements are met.

The recombinant DNA molecule carrying the desired structural gene under the control of the Enod2 regulatory region of soybean can be introduced into plant tissue by any means known to those skilled in the art. The technique used for a given plant species or specific type of plant tissue depends on the known successful techniques. As novel means are developed for the stable insertion of foreign genes into plant cells and for manipulating the modified cells, skilled artisans will be able to select from known means to achieve a desired result. Means for introducing recombinant DNA into plant tissue include, but are not limited to transformation (J. Paszkowski et al. (1984) EMBO J. 3:2717), electroporation (M. Fromm et al. (1985) Proc. Natl. Acad. Sci. USA 82:5824), microinjection (A. Crossway et al. (1986) Mol. Gen. Genet. 202:179), or T-DNA mediated transfer from Agrobacterium tumefaciens to

28

the plant tissue.   There appears to be no fundamental
limitation of T-DNA transformation to the natural host
range of _Agrobacterium_.   Successful T-DNA-mediated
transformation of monocots (G. Hooykaas-Van Slogteren _et_
_al_. (1984) Nature 311:763), gymnosperm (A. Dandekar _et_ _al_.
(1987) Biotechnol. 5:587) and algae (R. Ausich, EPO
Application 108,580) has been reported.   Representative
T-DNA vector systems are described in the following
references: G. An _et_ _al_. (1985) EMBO J. 4:277; L.
Herrera-Estrella _et_ _al_. (1983) Nature 303:209; L.
Herrera-Estrella _et_ _al_. (1983) EMBO J. 2:987; L.
Herrera-Estrella _et_ _al_. (1985) in _Plant_ _Genetic_
Engineering, New York: Cambridge University Press, p.63.
Once introduced into the plant tissue, the expression of
the structural gene may be assayed by any means known to
the art, and expression may be measured as mRNA transcribed
or as protein synthesized.   Techniques are known for the _in_
vitro culture of plant tissue, and in a number of cases,
for regeneration into whole plants.   Several methods are
known for the regeneration of soybean tissue.   Procedures
for transferring the introduced expression complex to
commercially useful cultivars are known to those skilled in
the art.

The skilled artisan can insert the Enod2 gene, or a
chimeric gene comprising the Enod2 regulatory region and a
downstream structural gene under the regulatory control of

29

said region, in an <u>Agrobacterium</u> <u>tumefaciens</u> T-DNA based
vector or an <u>Agrobacterium</u> <u>rhizogenes</u> T-DNA based shuttle
vector or a which will allow the transfer of the Enod2 gene
or the chimeric gene to soybean or to heterologous plant

5   hosts.   As will be readily apparent to those of ordinary
skill in the art, any plant-expressible gene can be incor-
porated in place of the Enod2 coding region of the expres-
sion complex using any naturally occurring or artificially
engineered restriction sites convenient for <u>in</u> <u>vitro</u>

10   manipulations.   The major consideration is that the
sequences at the junctions remain compatible with tran-
scriptional and translational functionality.   The final
steps for obtaining genetically modified plant tissue
include introducing the expression complex into plant

15   tissue, for example, by inserting the expression complex
into a T-DNA-containing vector, and transferring the
recombinant DNA to plant tissue wherein the modified T-DNA
becomes stably integrated as part of the genome.

The following examples are provided for illustrative

20   purposes only and are not intended to limit the scope of
the invention.   The examples utilize many techniques well
known and accessible to those skilled in the arts of
molecular biology, in the manipulation of recombinant DNA
in plant tissue, and in the culture and regeneration of

25   transformed plants.   Enzymes are obtained from commercial
sources and are used according to the vendors'

30

recommendations or other variations known in the art. Reagents, buffers and culture conditions are also known to the art. References containing standard molecular biological procedures include T. Maniatis et al. (1982) Molecular Cloning, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; R. Wu (ed.) (1979) Meth. Enzymol. 68; R. Wu et al. (eds.) (1983) Meth. Enzymol. 100 and 101: L. Grossman and K. Moldave (eds.) (1980) Meth. Enzymol. 65; J. Miller (ed.) (1972) Experiments in Molecular Genetics, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; Old and Primrose (1981) Principles of Gene Manipulation, University of California Press, Berkeley, California; R. Schlief and P. Wensink (1982) Practical Methods in Molecular Biology; Glover (ed.) (1985) DNA Cloning, Vols. I and II, IRL Press, Oxford, UK; Hames and Higgins (eds.) (1985) Nucleic Acid Hybridization, IRL Press, Ox ford, UK; Setlow and A. Hollaender (1979) Genetic Engineering: Principles and Methods, Vols. 1-4, Plenum Press, New York, which are expressly incorporated by reference herein. Abbreviations and nomenclature, where employed, are deemed standard in the field and commonly used in professional journals such as those cited herein.

Example 1:   Isolation of a cDNA clone homologous to the Enod2 gene

Soybean plants (Glycine max (L) Merr. cv. Williams) were cultured as described for pea plants (T. Bisseling et al. (1978) Biochim. Biophys. Acta 539:1-11) except at 28°C. At the time of sowing, the seeds were inoculated with Bradyrhizobium japonicum USDA110.   Nodules were excised from the roots with a scalpel. Nodules were frozen in liquid nitrogen and stored at -70°C until use.   Total RNA from nodules and roots was isolated as described   (F. Govers et al. (1985) EMBO J. 4:861-867).   Poly(A)$^+$ was obtained by oligo (dT)-cellulose chromatography and plasmid DNA was isolated by an alkaline lysis procedure (T. Maniatis et al. (1982), supra).

DNA complementary to poly(A)$^+$ RNA isolated from nodules from 21-day-old plants was synthesized with reverse transcriptase (Anglian Biotechnology, Essex, England) and second strand synthesis was performed under standard conditions (Maniatis et al. (1982), supra).   The double-stranded cDNA was treated with S1 nuclease (10 units per μg of ds cDNA) and fractionated on a 5-30% sucrose gradient (Beckman SW50 rotor, 47,000 rpm, 6 hr, 4°C). Double-stranded cDNA with a length of 500 bp or more was tailed with dC and then   annealed to PstI-cut oligo(dG)-tailed pBR322 (Boehringer Mannheim) in a 1:1 molar ratio.   The hybridized mixture was treated with DNA

ligase and used to transform <u>Escherichia</u> <u>coli</u> RR1 (Maniatis
<u>et</u> <u>al</u>. (1982), <u>supra</u>).

Individual transformants were picked, transferred to
96 well microtiter plates containing LB medium, 15%
glycerol, and 12.5 μg/ml tetracycline, and grown for 16 h
at 37°C. Two replicate filters were made on GeneScreen<u>Plus</u>
(New England Nuclear). After 16 hr of bacterial growth on
LB agar containing 12.5 μg/ml tetracycline, the filters
were prepared for hybridization according to the
manufacturer's instructions.

Probes for differential screening were prepared from
poly(A)$^+$ RNA isolated from segments of 5-day-old uninfected
roots and from nodules 10 days after inoculation. The
poly(A)$^+$ RNA was incubated as described for first strand
cDNA synthesis except that 10 μCi α-[^{32}P]-ATP (specific
activity = 3200 Ci/mMol; 1 Ci = 37 GBq; New England
Nuclear) was used. The filters were hybridized as
described (H.J. Franssen <u>et</u> <u>al</u>. (1987), <u>supra</u>).

Those clones which specifically hybridized to the
10-day-old nodule poly(A)+ RNA were designated Enod clones
because they represent early nodulin genes, which are
expressed in the early stages of nodule development. Nod
clones are those which represent nodulin genes expressed
later during nodule development. pEnod2, which had an

33

insert length of 1000 bp, was chosen for further characterization.

Example 2: Identification of pEnod2 as a gene encoding an early nodulin of about 75 kDa

5   The in vitro translation product of the mRNA homologous to pEnod2 was determined as described (Franssen et al. (1982), supra). The results showed that the pEnod2-encoded polypeptide had an apparent M_r of about 75,000, and an isoelectric point of about 6.5. In

10   accordance with the established nodulin nomenclature (A. van Kammen (1984) Plant Mol. Biol. Rep. 2:43-45), the identified polypeptide was called N-75. After in vitro translation of pEnod2-hybrid-selected mRNA in the presence of [3H]-leucine, a second polypeptide was found which was

15   slightly more basic than the polypeptide which co-migrated with that also labelled with [35S]-methionine.

The translation products of the Enod2 mRNAs of about 1200 nucleotides in length have a maximum coding capacity for a polypeptide of only about 45,000. This discrepancy

20   prompted an examination of the DNA sequence of the pEnod2 insert to determine if the deduced amino acid sequence could explain the anomalous size of the encoded proteins. Standard techniques were used for cloning into M13 and pUC vectors (J. Messing (1983) Meth. Enzymol. 101:20-78) and

25   for dideoxy (F. Sanger et al. (1977) Proc. Natl. Acad. Sci. USA 74:5463-5467; M.D. Biggen et al. (1983) Proc.

34

Natl. Acad. Sci. USA 80:3963-3965) and for Maxam-Gilbert

sequencing (A.M. Maxam and W. Gilbert (1980) Meth. Enzymol.

65:499-560). The DNA sequence data were stored and

analyzed with programs written by R. Staden (R. Staden

5   (1984) Nucleic Acids Res. 12:521-538) on a microVAX/VMS

computer.

Example 3: Isolation of genomic clones homologous to
            pEnod2

A soybean (Glycine max cv. Wayne) genomic library,

10   constructed as a Sau3A partial digest in lambdaCH35, was

obtained from R. Nagao and J. Key (University of Georgia,

Athens, GA). E. coli K802⁻ was used as the host for

lambdaCH35 clones. A large number of clones, 10^5,

representing 2x the soybean genome, were screened for the

15   presence of sequences hybridizing to the radiolabeled

pEnod2 probe.

CH6 and CH9 were the two genomic clones which

hybridized to the pEnod2 probe. Restriction site mapping

on CH6 and CH9 was performed using BamHI, EcoRI, and

20   HindIII (Figure 1). DNA was digested with restriction

enzymes and the fragments were separated by agarose gel

electrophoresis and subsequently blotted onto

GeneScreenPlus. Blots were probed with either complete

pEnod2 or PstI-HindIII clones prepared therefrom.

25   Restriction fragments containing sequences which hybridized

to the pEnod2 were subcloned into pBR322 and propagated in

E. coli HB101.   p4.5BE contained a 4.5 kb EcoRI fragment from CH6, and p10.2 contained a 10.2 HindIII-BamHI fragment from CH9.

Subsequently, portions of p4.5BE and p10.2 were
5   subcloned into pUC18 and pUC19 vectors and sequenced as described in Example 2. The DNA sequences of the portions of p4.5BE and p10.2 containing the Enod2 genes are displayed in Tables 1 and 2.   The coding regions and the deduced amino acid sequences of both genes are shown.

10   Example 4:   Sequence analysis of the Enod2a and Enod2b
                    genes of soybean

Standard techniques, as described above, were used for the sequencing of the Enod2a and Enod2b genomic sequences.   The coding region of each of these genes is an
15   uninterrupted sequence of 930 bp.   Table 1 gives the DNA sequence of the coding region of the Enod2a gene along with about 1650 bp of 5' flanking sequence and about 360 bp of 3' flanking sequence.   The coding region and about 600 of 5'   flanking sequence of the Enod2b gene is almost
20   identical in sequence to that of the Enod2a gene as shown in Table 2; a total of about 2450 of 5' flanking sequence and about 470 bp of 3' flanking sequence of the Enod2b gene are also presented in Table 2.   It was noted that the two genes were 100% homologous over the coding regions, and
25   almost 100% homologous in the approximately 600 bp of 5' flanking DNA extending to a Sau3A site at positions 1048 in

36

Enod2a and 1852 in Enod2b, and in the 3' flanking DNA that has been sequenced.

Analysis of the sequence of the cDNA clone pEnod2 and the sequences revealed that there were two open reading
5 frames (ORF1 and ORF2) of similar length; both are noted in Tables 1 and 2. The anomalous migration in SDS-polyacrylamide gel electrophoresis experiments led, in part, to the conclusion that the ORF1 is the actual coding sequence of the Enod2 genes encoding N-75. The polypeptide
10 encoded by ORF1 is rich in proline, and proline-rich polypeptides are known to exhibit aberrant behavior during SDS-polyacrylamide gel electrophoresis (J.W. Freytag et al. (1979), supra). The second line of reasoning was that one of the hybrid-selected translation products was devoid of
15 methionine; ORF1 has only one methionine codon (at the translation start) while the alternate ORF1 contained about 20 methionine codons, and therefore its translation product should have been labelled readily with $[^{35}S]$-methionine.

37

Table 1: Nucleotide Sequence of Enod2a Genomic Clone

```
                              GGATCCTTACACAGGCCAGACATCCCCAAGTTCTCA   36
                              BamHI
AATAAGACAAATTTGGTTGTTCTTTTCTTAATATTTCACAGGGAGATGTTCTGTCTTTTG   96

ATTTGGGGATTTCATTTAGCACATAACAAACAGTTAACAAAATTTCGCCCCACCAAAAAG  156

ATGTTGCACTAGAACTCAACATAGTAGCTACAACTAATTCTGTAAAAGTTCTGTTCTTTC  216

TTTCAGCTTTACCGTTCATTTCAGGTGAATATGGAGCAGTTGTTTCATGTATGATTCCAT  276

GCAAATTATAAAACTCATTAAACAAACTGGAATCATACTCTGTGCCTCTATCACTTCGAA  336

GTTTCTTAATTTTCTTATTGAATTGATTTTCAATTTCTGTTACAAATAACTTAAACATGT  396

CAAGCCCTTCACTTTTATTTTTCATAAGATATACATATATATATAATCAGAGCAGTCATC  456

AATAAAGTGATAAAATATCGTTTTCCATTTCTGGTCAACGTTCCATCAAATTCACATAT   516

ATCAGAATGTATTAAATCCAATGGCTCAGATTCTTTAACTACTGATTTTTGTGATTTTT   576

AGTTATTTTAGATTGACTGCAAAAAACACACTTTTCAAAGTGATTTGAAGATAGCTTTGG  636

AATAAAACCTAAGTTACTCATATTAGATATGCAACGACTATTTATATGACAAAGTCTAGA  696

ATGCCAGAATTAAAATCACACAGCATGTAAGCAGAAGGAGAAACTTTATTAATATCAAGA  756

TTCAATTTGAACATGCCATCAGTGGCGTACCCTTCCCTACAAATACCCATTCTTGGTC    816

AAAGTAAATAAATCTGCACCTATGGTCTGAGTAAACCCAGCCTTGTTTAAAAGAAAACCA  876

GAAACCAGATTCTTTCTCATCTCTGGAGTATGCATCACATCTTTGAGAATCAAAGTCTTT  936

CCAGAGGTAAACTTCAGTTCAACATCTCCAGTTCTAGCAACAGTAGTGGTGTGGGAATCA  996
```

CCCAACAACACTTTCTTATTTTCAACATTCGTGTATGTTTTAAACATAGCATGATCTTTA 1056
<div style="text-align:right">Sau3A</div>

TACTTGTATTTTTTTTTTGTTTTAGTTTCTATACTTAAAAATTCTGTTTTATTATTTTTA 1116

CGCCTTAGTTTTCTAGCAATCTAAAACTGATATAAAATAGAAGTATAACGACTAAAACAT 1176

AAAAAAAAAAATTGTATAAAAATAAAGCATATAGCTTTCATTCATATATAAGAACTAA 1236

ACTGAAATACCAGTGTAAGTATAAGAACTAATCGATAAATTAAGCCAAATTAAGGGTACA 1296

TATTATTTTTAAGAAAATTAGGCCGGGTATATATTTTTAAAAAGGACTATACACTATGTG 1356

ACGATAGAAATAATAGGTATGTAGATGTATGTTAAGTATTTTCTAATGTGTTTTTTACTT 1416

TCTCTATCACACTTGTTATTTTCTCACTATTTTTTTCTCTTGTTTCTCTGTTATTTTCAC 1476

TCTAAAACTGGAGTAATATGTTTATGACTACAACACATTTTGACATGACTTAGGATTAAC 1536
<div style="text-align:right">Hind3</div>
ATATATTATGATAAAATAACTAAAGATTGATAACCTTGATAGAAAGCTTCTCATGTCTC 1596
<div style="text-align:left">TS ± 20</div>

CTCTCCCTATAAGTAGTTTCCCATTGTTATCACTTTTCATCAGCACAAGCTAAGACATGA 1656
<div style="text-align:right">ORF1 M</div>

CTTCTGTACTACACTACTCACTCCTGCTGCTCCTGCTTGGAGTGGTGATTCTCACCACTC 1716
T S V L H Y S L L L L L L G V V I L T T

CAGTGCTAGCTAATTTGAAGCCACGCTTCTTCTATGAGCCTCCTCCAATTGAGAAACCCC 1776
P V L A N L K P R F F Y E P P P I E K P
<div style="text-align:center">ORF2   M S L L Q L R N P</div>

CCACCTATGAACCTCCACCATTTTATAAGCCCCCATACTACCCACCACCAGTGCACCACC 1836
P T Y E P P P F Y K P P Y Y P P P V H H
P P M N L H H F I S P H T T H H Q C T T

CTCCACCAGAGTACCAACCACCCATGAAAAAACACCACCTGAGTATCTACCTCCTCCTC 1896
P P P E Y Q P P H E K T P P E Y L P P
L H Q S T N H P M K K H H L S I Y L L

ATGAGAAACCACCACCAGAATACCTACCTCCTCATGAGAAACCGCCACCAGAATACCAAC 1956
H E K P P P E Y L P P H E K P P P E Y Q
M R N H H Q N T Y L L M R N R H Q N T N

```
CTCCTCATGAGAAACCACCCCATGAGAATCCACCACCGGAGCACCAACCACCTCATGAGA 2016
P  P  H  E  K  P  P  H  E  N  P  P  P  E  H  Q  P  P  H  E
L  L  M  R  N  H  P  M  R  I  H  H  R  S  T  N  H  L  M  R

AGCCACCAGAGCACCAACCACCTCATGAGAAGCCACCACCAGAGTATGAACCACCTCATG 2076
K  P  P  E  H  Q  P  P  H  E  K  P  P  P  E  Y  E  P  P  H
S  H  Q  S  T  N  H  L  M  R  S  H  H  Q  S  M  N  H  L  M

AGAAACCACCACCAGAATACCAACCACCTCATGAGAAGCCACCACCAGAATACCAACCAC 2136
E  K  P  P  P  E  Y  Q  P  P  H  E  K  P  P  P  E  Y  Q  P
R  N  H  H  Q  N  T  N  H  L  M  R  S  H  H  Q  N  T  N  H

CTCATGAGAAACCACCACCAGAATACCAACCACCTCATGAGAAGCCACCACCAGAGCACC 2196
P  H  E  K  P  P  P  E  Y  Q  P  P  H  E  K  P  P  P  E  H
L  M  R  N  H  H  Q  N  T  N  H  L  M  R  S  H  H  Q  S  T

AACCACCTCATGAGAAGCCACCAGAGCACCAGCCACCTCATGAGAAGCCACCACCAGAGT 2256
Q  P  P  H  E  K  P  P  P  E  H  Q  P  P  H  E  K  P  P  P  E
N  H  L  M  R  S  H  Q  S  T  S  H  L  M  R  S  H  H  Q  S

ATCAACCACCTCATGAGAAACCACCACCAGAATACCAACCTCCTCAAGAAAGCCACCAC 2316
Y  Q  P  P  H  E  K  P  P  P  E  Y  Q  P  P  Q  E  K  P  P
I  N  H  L  M  R  N  H  H  Q  N  T  N  L  L  K  K  S  H  H

ATGAAAACCACCGCCAGAATACCAACCTCCTCATGAAAAGCCACCACCAGAACACCAAC 2376
H  E  K  P  P  P  E  Y  Q  P  P  H  E  K  P  P  P  E  H  Q
M  K  N  H  R  Q  N  T  N  L  L  M  K  S  H  H  Q  N  T  N

CTCCCCATGAAAAGCCACCACCAGTGTACCCACCCCCTTATGAGAAACCACCACCAGTGT 2436
P  P  H  E  K  P  P  P  V  Y  P  P  P  Y  E  K  P  P  P  V
L  P  M  K  S  H  H  Q  C  T  H  P  L  M  R  N  H  H  Q  C

ATGAACCCCCTTATGAGAAGCCACCCCCAGTAGTGTATCCACCTCCTCATGAGAAACCAC 2496
Y  E  P  P  Y  E  K  P  P  P  V  V  Y  P  P  P  H  E  K  P
M  N  P  L  M  R  S  H  P  Q  *  Y  P  P  P  H  E  K  P

CCATTTATGAGCCACCGCCATTGGAGAAGCCACCGGTCTACAATCCCCCACCTTATGGCC 2556
P  I  Y  E  P  P  P  L  E  K  P  P  V  Y  N  P  P  P  Y  G

GCTATCCACCATCCAAGAAAAACTAATAACCACTTGCCTGCGTCACATGTTTGGTCTAC 2616
R  Y  P  P  S  K  K  N  *

TCAAACTTAGACCTGCCCTTTGTCATATAAAGCTTTCTGTTTCTGTTTAAGATCTCAAGT 2676
                                Hind3

ACAATATGTCCCTTCTGCATGCACTACTTCTTCAAAATAAAGGCTTTATGCCTATGTATA 2736

ATACTCTACTTTAATTCTCCTTTCACCATCGATATTGTAATGTCAACTACTAGTGTGGGT 2796

TTATCTATGGCTATAATAAGTTTTTCTTTGTGTTTACTTATGAGTCTTTGTTTTTAATTG 2856
```

42

CATGCTAAAAATTGGCAAAAACATATATAATTCTGTTCGTACATGTTTTATTTTATGAAC 2916

TTCATAAGTACCGGTAAAGCAATGATAATGTGTAAAGTTGCTTGGTCTATATATATGTTT 2976

AAATACACATATCTCTAAACCGTCAATGAGAAATACTCTCTGTACCTGTTTATTCAACTT 3036

GGAAAACTAAACCACATAATAAAC
                                                          3060

Table 2: Nucleotide Sequence of Enod2b Genomic Clone

```
AAGCTTGACAAAAGATAAATGCTTTGTGGGGTGGCGTAGCGTCTTTATGCAGCAATGGTT     60
Hind3

TATGTAATTTATGTAATGGGGTGGTCACTCCTAGTGACTGTCCTCTGTGTTATGATTAAT    120

GAAATGTTTTGCTTTTTCGAAAAGAACAAAAAATCCTTAAGTTCACCCCATTTGTAAATA    180

GTCTCTTACATTGAATTGGGGTTGAATTATTAAAGAAGAAATCTCAACTACTTATTTATT    240

TTAAATTTCAATCATTTATTAGTTTAATTTTTATAAATCACTTTTCTAAATATTAAAATA    300

TAATAAAACTCTTCTAAAAACATAATAAAATTAATAACTAAAATAAATAAATTATTTTTT    360

ATTGGTATTTATTTTTGTTTTTTTTTTTTTCTAAATTCATATTCTTTTACTTATGTTTTAA    420

TAGACAAAAACTGATTTGTAACXXXXXXCATGTATAGAAAACTATTCCTTTAACCTATA    480

AAAAACTATCATTAAAATATTTTTTAAGATAATTATTATAAAAATCAACAAACTTATTAA    540

TAATATATGATTCAATAATAATATATAAAATCTTTGCATCTAACATAAATTATAATAATA    600

TTACAATTTTTTCCTTTAAATCAATTTTACATTTTAAAAAATCAAATTAAATTCATATCC    660

GACTATTGCTGCGCATGATAGGCTCTAAAAGACCATCCCATTCACATATTAATATCTTAT    720

TCAACGTTAATCTGTGTTCTGTTAGATTCCAAAGATTCCAGTGAATAGTGATGGCTAAGA    780

ACAGTTTCTTGACCTTTCGCTAACAGCAAGCCTACCTATACAAGCTCCAATTATTTTCT    840

TTTTTGAGGATTGCTCCATTTATTXXCCGACAAAACATACATGCATCTAAATGTGGCAGC    900

ATGCTAAAGTTTTGGTGAGGCTATAGTAAAATATGAAATAAAGATTTGAAGTTTCAGCCC    960

AATATAAAAAAAAAATTAATTCCTTCTGAAATGAAAAGAGTATCAAAGAAGATATAATCA   1020
```

GTAAAATCTTTTTCATAAGCATTGATCTGGATACATCAACTTTGATGCGTTGGAAATACT 1080

GTGCTCAAGTTTGACAGCAATTCTTGGAATTTTTTCGCCACAACAGAAGCTCCAGACGAT 1140

TATGATTTATGACCTTATATGATGTTAGTTACGTGAAAGTAATTAGAATCGCATTTGCTA 1200

ACTATTAGCAATTTTTTTTTTTAAGCTAATGCAAGTGACAGAATCTTAGGTCTCTATAAT 1260

TTGAACCTGTGGCGGTGGAACTCGTACTTCATGTGCTGAAAAGAACTTGATATTTTTTTA 1320

AGGGAAATAATATATATCAATGCTCCTAAGTCCTAAACTTTATCTTCTTTGGCAGCTAAA 1380

TTTACTTTAAAAAGAAATAAGATTAAATAACTTTTTCTTACAAGAAAATATATTTAATTA 1440

TTAATTGTTAAGTTTAACGTCTTTTTATACATTTATTTGTTTTAAATTCCAGTCATCTTT 1500

TTAACATAATTCCAATCATTTATTAGTTTTACTTTTATAAACAATAAAACATAATTAATT 1560

TTCAGATTAAAAAATAGATAGAAGTTTTTTTAATTGTTTTTTATTATCAAATTTCAATTTT 1620

AACATATTTTATAATAGATAAAATGAATTGTAACAAATTAATGATTGACCTTATAGATAA 1680

GTAATTAGCCAACAACTTTTTTAGTATTAAATTGATAGAAAAATTAAGCTATATTTGGG 1740

GGGGGGGGGGGTCAAGTTTAATGAAGTTAAAGTTCATTGAATATATTTGTAAAAAAGAT 1800

AAAGGGTTTAAGGTCTAATAGAGATAATATTTAAGGACTTAATTAATTATTTGATCTTTA 1360

SauJa

TACTTGTATTTTTTTTTGTTTAGTTTCTATACTTAAAAATTCTGTTTTATTATTTTTA 1920

CGCCTTAGTTTTCTAGCAATCTAAAACTGATATAAAATAGAAGTATAACGACTAAAACAT 1980

AAAAAAAAAAAATTGTATAAAAAATAAAGCATATAGCTTTCATTCATATATAAGAACTAA 2040

ACTGAAATACCAGTGTAAGTATAAGAACTAATCGATAAATTAAGCCAAATTAAGGGTACA 2100

```
TATTATTTTTAAGAAAATTAGGCCCGGGTATATATTTTTAAAAAGGACTATACACTATGTG 2160

ACGATAGAAATAATAGGTATGTAGATGTATGTTAAGTATTTTCTAATGTGTTTTTTACTT 2220

TCTCTATCACACTTGTTATTTTCTCACTATTTTTTTCTCTTGTTTCTCTGTTATTTTCAC 2280

TCTAAAACTGGAGTAATATGTTTATGACTACAACACATTTTGACATGACTTAGGATTAAC 2340

ATATATTATGATAAAATAACTAAAGATTGATAACCTTGATAGAAGCTTCTCATGTCTCCT 2400
    TS ± 20                                  Hind3

CTCCCTATAAGTAGTTTCCCATTGTTATCACTTTTCATCAGCACAAGCTAAGACATGACT 2460
                                                      ORF1   M   T

TCTGTACTACACTACTCACTCCTGCTGCTCCTGCTTGGAGTGGTGATTCTCACCACTCCA 2520
 S   V   L   H   Y   S   L   L   L   L   L   L   G   V   V   I   L   T   T   P

GTGCTAGCTAATTTGAAGCCACGGCTTCTTCTATGAGCCTCCTCCAATTGAGAAACCCCCC 2580
 V   L   A   N   L   K   P   R   F   F   Y   E   P   P   P   I   E   K   P   P
                             ORF2   M   S   L   L   Q   L   R   N   P   P

ACCTATGAACCTCCACCATTTTATAAGCCCCCATACTACCCACCACCAGTGCACCACCCT 2640
 T   Y   E   P   P   P   F   Y   K   P   P   P   Y   Y   P   P   P   V   H   H   P
 P   M   N   L   H   H   P   I   S   P   H   T   T   H   H   Q   C   T   T   L

CCACCAGAGTACCAACCACCCCATGAAAAAACACCACCTGAGTATCTACCTCCTCCTCAT 2700
 P   P   E   Y   Q   P   P   H   E   K   T   P   P   E   Y   L   P   P   P   H
 H   Q   S   T   N   H   P   M   K   K   H   H   L   S   I   Y   L   L   L   M

GAGAAACCACCACCAGAATACCTACCTCCTCATGAGAAACCGCCACCAGAATACCAACCT 2760
 E   K   P   P   P   E   Y   L   P   P   H   E   K   P   P   P   E   Y   Q   P
 R   N   H   H   Q   N   T   Y   L   L   M   R   N   R   H   Q   N   T   N   L

CCTCATGAGAAACCACCCCATGAGAATCCACCACCGGAGCACCAACCACCTCATGAGAAG 2820
 P   H   E   K   P   P   H   E   N   P   P   P   E   H   Q   P   P   H   E   K
 L   M   R   N   H   P   M   R   I   H   H   R   S   T   N   H   L   M   R   S

CCACCAGAGCACCAACCACCTCATGAGAAGCCACCACCAGAGTATGAACCACCTCATGAG 2880
 P   P   E   H   Q   P   P   H   E   K   P   P   P   E   Y   E   P   P   H   E
 H   Q   S   T   N   H   L   M   R   S   H   H   Q   S   M   N   H   L   M   R

AAACCACCACCAGAATACCAACCACCTCATGAGAAGCCACCACCAGAATACCAACCACCT 2940
 K   P   P   P   E   Y   Q   P   P   H   E   K   P   P   P   E   Y   Q   P   P
 N   H   H   Q   N   T   N   H   L   M   R   S   H   H   Q   N   T   N   H   L

CATGAGAAACCACCACCAGAATACCAACCACCTCATGAGAAGCCACCACCAGAGCACCAA 3000
 H   E   K   P   P   P   E   Y   Q   P   P   H   E   K   P   P   P   E   H   Q
 M   R   N   H   H   Q   N   T   N   H   L   M   R   S   H   H   Q   S   T   N
```

```
CCACCTCATGAGAAGCCACCAGAGCACCAGCCACCTCATGAGAAGCCACCACCAGAGTAT 3060
 P  P  H  E  K  P  P  E  H  Q  P  P  H  E  K  P  P  P  E  Y
   H  L  M  R  S  H  Q  S  T  S  H  L  M  R  S  H  H  Q  S  I

CAACCACCTCATGAGAAACCACCACCAGAATACCAACCTCCTCAAGAAAAGCCACCACAT 3120
 Q  P  P  H  E  K  P  P  P  E  Y  Q  P  P  Q  E  K  P  P  H
   N  H  L  M  R  N  H  H  Q  N  T  N  L  L  K  K  S  H  H  M

GAAAAACCACCGCCAGAATACCAACCTCCTCATGAAAGCCACCACCAGAACACCAACCT 3180
 E  K  P  P  P  E  Y  Q  P  P  H  E  K  P  P  P  E  H  Q  P
   K  N  H  R  Q  N  T  N  L  L  M  K  S  H  H  Q  N  T  N  L

CCCCATGAAAAGCCACCACCAGTGTACCCACCCCCTTATGAGAAACCACCACCAGTGTAT 3240
 P  H  E  K  P  P  P  V  Y  P  P  P  Y  E  K  P  P  P  V  Y
   P  M  K  S  H  H  Q  C  T  H  P  L  M  R  N  H  H  Q  C  M

GAACCCCCTTATGAGAAGCCACCCCCAGTAGTGTATCCACCTCCTCATGAGAAACCACCC 3300
 E  P  P  Y  E  K  P  P  P  V  V  Y  P  P  P  H  E  K  P  P
   N  P  L  M  R  S  H  P  Q  *  V  V  Y  P  P  P  H  E  K  P  P

ATTTATGAGCCACCGCCATTGGAGAAGCCACCGGTCTACAATCCCCCACCTTATGGCCGC 3360
 I  Y  E  P  P  P  L  E  K  P  P  V  Y  N  P  P  P  Y  G  R

TATCCACCATCCAAGAAAAACTAATAACCACTTGCCTGCGTCACATGTTTTGGTCTACTC 3420
 Y  P  P  S  K  K  N  *

AAACTTAGACCTGCCCTTTGTCATATAAAGCTTTCTGTTTCTGTTTAAGATCTCAAGTAC 3480
                         Hind3

AATATGTCCCTTCTGCATGCACTACTTCTTCAAAATAAAGGCTTTATGCCTATGTATAAT 3540

ACTCTACTTTAATTCTCCTTTCACCATCGATATTGTAATGTCAACTACTAGTGTGGGTTT 3600

ATCTATGGCTATAATAAGTTTTTCTTTGTGTTTACTTATGAGTCTTTGTTTTTAATTGCA 3660

TGCTAAAAATTGGCAAAAACATATATAATTCTGTTCGTACATGTTTTATTTTATGAACTT 3720

CATAAGTACCGGTAAAGCAATGATAATGTGTAAAGTTGCTTGGTCTATATATATGTTTAA 3780

ATACACATATCTCTAAACCTGTCAATGAGAAATACTCTCTTGTACCTTGTTTATTCAACT 3840

TGGGAGACTAAACCTA
                                                      3856
```